DOI:
10.1038/ng.3979

Document Version
Peer reviewed version

Predicting causal variants affecting expression using whole genome sequence and RNA-seq from multiple human tissues

Andrew Anand Brown[1,2,3,4,*], Ana Viñuela[1,2,3], Olivier Delaneau[1,2,3], Tim D Spector[1,2,3], Kerrin S Small[5], Emmanouil T Dermitzakis[1,2,3,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland.

[2]Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland.

[3]Swiss Institute of Bioinformatics, Geneva, Switzerland.

[4]NORMENT, KG Jebsen Centre for Psychosis Research, Oslo University Hospital, Oslo, Norway.

[5]Department of Twin Research and Genetic Epidemiology, King's College London, St Thomas' Campus, London, United Kingdom.

[*]Corresponding authors. Email: andrew.brown@unige.ch and emmanouil.dermitzakis@unige.ch

Genetic association mapping produces statistical links between phenotypes and genomic regions, but identifying causal variants remains difficult. Complete knowledge of all genetic variants, as provided by whole genome sequence (WGS), will help, but is financially prohibitive for well powered GWAS studies. We performed eQTL mapping using WGS and RNA-seq, and

showed lead eQTL variants called using WGS are more likely to be causal. We derived properties of causal variants using simulations, and used these to propose a method for implicating likely causal SNPs. We estimate that 25% - 70% of causal variants lie in open chromatin regions, depending on tissue and experiment. Finally, we identify a set of high confidence causal variants and show they are more enriched in GWAS associations than other eQTLs. Of these, we find 65 associations with GWAS traits, giving examples where the gene implicated by expression has been functionally validated as relevant for complex traits.

Genome-wide associations studies (GWAS) have uncovered 1,000s of genetic associations between regions of the genome and complex traits[1], but moving from the association to identifying the mechanism behind it has proven complicated[2]. Statistical associations between traits and genomic regions indicate a variant with a causal effect on the trait, as reverse causation or unmeasured confounders modifying DNA can be ruled out (we interpret causal effect in the probabilistic sense, where a direct intervention modifying one factor has consequences on another). A first step for understanding the mechanism would be to identify the exact variant, as exact localisation would allow exploration as to which transcription factor binding sites and regulatory elements are affected. This, however, is complicated by the fact that most loci tested in GWAS studies are not directly measured, but instead imperfectly imputed[3]. Whole-genome sequence (WGS) data does directly ascertain all genotype calls, but despite falling costs is still very expensive on the sample sizes of modern GWAS studies (Supplementary Table 1). In contrast, eQTL studies linking variants and gene expression have discovered 1,000s of associations using few hundreds of samples, a scale at which collecting whole genome sequence data is feasible[4].

We describe analysis combining for the first time two previously published datasets derived from individuals in the TwinsUK cohort: RNA-seq from four tissues[5,6] and WGS from the UK10K project[7] (previously, gene expression quantified using micro-arrays[8] has been combined with the same whole genome sequence dataset for specific look-ups of GWAS associations[9,10]). We explore the properties of causal variants using simulations, and propose the CaVEMaN method (Causal Variant Evidence Mapping using Non-parametric resampling) to estimate the probability that the variant most associated with the expression trait is causal for that association. We use this to produce a robust set of likely causal SNPs; this could be an important resource for developing methods to call personalised regulatory variants from whole-genome sequence and sequence annotations.

With whole genome sequence, genotypes are directly measured at a far more sites than are available on current genotyping chip arrays (although sites on a genotyping chip should be measured with more accuracy). The 1000 Genomes project estimate they observe > 99% of SNPs with minor allele frequency greater than 1%. For low coverage sequencing and genotyping arrays, imputation methods are frequently used to impute better quality calls at sites with no coverage on the arrays and low or no coverage with sequence data. The degree to which, if at all, information at more sites from sequence reduces imputation noise and increases power to map eQTLs is currently unknown. For a simple comparison, we mapped independent eQTLs within 1Mb of the transcription start site for protein coding genes and lincRNAs in four tissues (fat, lymphoblastoid cell lines (LCLs), skin and whole blood) using individuals for which expression, sequence and genotype array data were all available (N from 242 (whole blood) to 506 (LCLs)). We identify 27,659 independent autosomal eQTLs affecting 11,865 genes using whole genome sequence (8,690,715 variants), and 26,351 affecting 11,642

genes using genotypes called from arrays and imputed into the 1000 Genomes Project Phase 1 reference panel (6,263,243 variants) (Figure 1, an analysis of all individuals with expression and WGS data (N from 246-523) and including the X chromosome found 28,141 eQTLs affecting 12,243 genes). This means just a 3.7% increase in discovered eQTLs using WGS; balanced against at least a ten-fold increase in cost of collecting the data, it does not seem a worthwhile exercise yet. This demonstrates the ability of imputation approaches to accurately assay common variation, particularly since the denser genotyping arrays and larger reference panels now available would reduce and possibly even remove this difference (more details on imputation accuracy is available in the Online Methods).

We frequently observe that the lead eQTL variant (LEV, the variant most associated with the trait) differs between the two datasets. As genotypic uncertainty should be reduced for WGS, we expect the WGS LEVs to be the causal variant more frequently than LEVs from genotype arrays. To test this hypothesis, we looked for enrichment of WGS-derived LEVs relative to array-genotype-derived in biochemically active regions of the genome. Indeed, for 30 out of 31 experiments carried out by the Roadmap Epigenomics consortium[11] in relevant tissues, we see significant enrichment of sequence LEVs compared to genotype LEVs falling in DNase1 hypersensitivity sites (DHS) (Odds ratio, 1.17-1.40, Figure 2). From this we infer that the LEVs called with our sequence are more likely the causal variant.

To better understand properties of causal variants we simulated expression datasets where the causal variant is known, with properties matched to those of the LEVs from the original eQTL mapping with sequence genotypes (considering effect size, distance to the transcription start site and minor allele frequency). Repeating the eQTL mapping on these simulated datasets, we found that in 45% of cases the causal variant was the LEV. This number was

consistent across tissues, despite sample size and power to map eQTLs being much reduced for whole blood (Supplementary Figure 1). This number is also similar to that obtained from the analysis of the Geuvadis data (55%), using a different methodology. We also see a rapid decline forlower ranked candidate variants, the 10th most associated SNP is causal in only 1% of cases.

Our simulations show that across all genes, the LEV is a strong candidate for the causal variant. However, for specific LEVs, causality will depend on the linkage disequilibrium structure around the true causal variant and phenotypic uncertainty in expression of the particular gene. For these reasons we developed the CaVEMaN method, which uses bootstrap methods similar to those previously proposed by others[12,13] to estimate the probability that the LEV is the causal variant (see Online Methods for methodological details).

We applied the CaVEMaN method to all four tissues and the Geuvadis LCL RNA-seq data (N = 445, results in Supplementary Data Set 1). The distributions of probabilities that LEVs are causal are similar across tissues and studies (Figure 3). For 7.5% of the eQTLs the LEV has P > 0.8 of being the causal variant, we refer to these as High Confidence Causal Variants (HCCVs). For comparison, we applied the CAVIAR method[14] to the largest dataset (TwinsUK LCLs), and dap-g[15] to simulated data (full details in the Online Methods).

To understand more about the relationship between causal regulatory variation and active genomic regions found by ChIP-seq in single individuals, we integrated our causal probabilities with DHSs from the Roadmap Epigenomics consortium. Figure 4 shows a simple linear relationship between the causal probability of the LEV and the probability that the LEV is located in a DHS (though low probability blood eQTLs (P < 0.25) are found less often in DHSs than expected by the linear model, possibly due to these LEVs being less reliable due to the

117 smaller sample size). We exploit the linear relationship to estimate the proportion of
118 regulatory variants with causal probability 1 that lie within DHS identified by particular
119 experiments. Figure 5 shows that for all tissues except blood, only a minority of regulatory
120 variants lie within DHS called by specific experiments. Blood eQTLs, discovered in a smaller
121 sample size than the other tissues, have larger effect sizes and thus are more likely to affect
122 promoter activity, a possible explanation for the observed greater enrichment. When
123 CaVEMaN is applied to larger eQTL datasets, with power to discover eQTLs with more subtle
124 effects, it is possible the proportion of causal regulatory variants in DHSs will be even lower,
125 implying limited utility of regulatory annotations for interpretation of enhancer and weaker
126 regulatory variants.

127 It is widely known that associations with whole organism traits, as discovered by GWAS, are
128 enriched in eQTLs[20]; by defining a set of eQTLs where the causal variant is known with high
129 probability, these could show greater enrichment (a shared GWAS-eQTL signal would not be
130 diluted by linkage). In addition, by providing both a mediating gene and a variant causative for
131 the expression signal, these results could provide a more mechanistic understanding of GWAS
132 signal. We extracted P values for association for all of the LEVs from 16 GWAS studies with
133 publicly available summary statistics (see Online Methods) and saw greater enrichment of
134 small P values for HCCVs compared to all other eQTLs (π_1 = 16.2 compared with π_1 = 14.0,
135 estimated using qvalue[21]). Greater enrichment was also observed when considering the
136 proportion of shared signals between GWAS associations with $P < 5\times10^{-8}$ listed in the NHGRI
137 catalogue and eQTLs falling in the same recombination hotspot (16.0% of proximal HCCVs and
138 GWAS associations were shared, 2.49% for all other eQTLs, estimated using the Regulatory
139 Concordance method[22,23] (RTC)). We also found Bonferroni significant GWAS associations

between 53 HCCVs and 65 GWAS traits ($P < 3\times10^{-6}$, Figure 6, Supplementary Data Set 2). Applying the coloc method to test whether the eQTL and GWAS trait are affected by the same causal variant[24], we observe 18 cases where there is strong evidence of common genetic effects (coloc probability > 0.95) and 29 cases with at least moderate evidence (coloc probability > 0.7).

Given these examples of variants with highly confident causal effects on expression and statistical associations with GWAS traits, functional evidence connecting the expression of the gene with the trait would also implicate a causal link between variant and trait. For example, a HCCV (rs10274367, all rs IDs are as defined in dbSNP, build 148, GRCh37) associated with *GPER1* is also associated with levels of high-density lipoprotein (HDL) cholesterol (coloc estimate of shared causal variant = 0.999). Female knock-out mice for the gene show a decrease in HDL levels[25]. We also found rs1805081 to be a HCCV for *NPC1*, and the lead associated variant with BMI in a large GWAS study[26] (coloc probability = 0.722). Heterozygous mouse models (Npc1+/-), where the gene is expressed at half normal levels, observe large weight gain on high fat diets but not on low fat diets[27,28], and higher levels of *NPC1* in human adipose tissue normalise after bariatric surgery and behavioural modification[29]. In this example, the expression of *NPC1* is modified by rs1805081 and hypothesised to be a response to changes in BMI. Expression changes in *NPC1* seem to be part of a compensatory mechanism to modify the weight gain due to dietary excesses and the result of diet-by-genotype interactions. Finally, we observe rs4702 as a HCCV for the *FURIN* gene in our analysis and it was the lead variant in the GWAS study of schizophrenia[30], coloc probability = 0.999). Altering expression of *FURIN* was seen to produce neuro-anatomical deficits in zebrafish and abnormal neural migration in human induced pluripotent stem cells[31].

This paper has produced a method for looking for causal variants for expression. However a HCCV associated with a GWAS trait does not necessarily mean that they shares a common causal variant, or the causal mechanism acts in the tissue under study. However, combining fine-mapping using CaVEMaN with co-localisation methods that formally test whether genetic variants between traits are shared[22,24] and methods which aim to predict causal tissues[23,32] could pinpoint precise variants, genes and tissues underlying GWAS traits. Also, methods for fine-mapping and for testing for co-localisation share common features. Similarly to how fine-mapping method (CAVIAR[14]) was extended to test for co-localisation (eCaviar[32]), CaVEMaN could also be extended to test for co-localisation.

In summary, we have produced a method to estimate the probability that the lead eQTL variant is the causal variant. We have used this method to estimate the effectiveness of ChIP-seq experiments from a single individual in predicting regions which harbour regulatory variation, and also to suggest variants which may be causal for GWAS associations. This method could also be applied to GWAS studies, learning candidate causal variants for whole organism traits. Pinpointing the causal variant in such studies will facilitate the integration of these association signals with mechanistic regulatory interactions and likely upstream regulators, and also allow the development of interpretation methods from genome sequence alone once a large number of representative causal variants have been discovered.

Acknowledgments

We would like to thank N. Lykoskoufis for his help with the enrichment analysis. This work has been supported by grants from the NIH-NIMH (GTEx), European Commission (Direct project), Louis Jeanet Foundation, Swiss National Science Foundation and SystemsX. The TwinsUK study was funded by the Wellcome Trust; European Community's Seventh Framework Programme

(FP7/2007-2013). The study also receives support from the National Institute for Health Research (NIHR)- funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. SNP genotyping was performed by The Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR. This study makes use of the data generated by the UK10K Consortium. Funding for UK10K was provided by the Wellcome Trust under award WT091310. A full list of the investigators who contributed to the generation of the data is available at www.UK10K.org. This research was supported by grants from the European Research Council. Computation was performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

Author Contributions

A.A.B. and E.T.D. designed the study. A.A.B. ran the analyses. A.V. provided interpretation of the results. A.A.B., A.V. and E.T.D. wrote the manuscript. O.D. made suggestions for the methods. K.S.S. and T.D.S. contributed data.

Competing Financial Interests Statement

There are no competing financial interests.

References

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).
2. Spain, S.L. & Barrett, J.C. Strategies for fine-mapping complex traits. *Human molecular genetics* **24**, R111-R119 (2015).
3. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
4. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
5. Brown, A.A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* **3**, e01381 (2014).
6. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet* **47**, 88-91 (2015).

7. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
8. Grundberg, E. *et al*. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).
9. Timpson, N.J. *et al*. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature communications* **5**(2014).
10. Iotchkova, V. *et al*. Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nature genetics* **48**, 1303-1312 (2016).
11. Kundaje, A. *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
12. Lebreton, C.M. & Visscher, P.M. Empirical nonparametric bootstrap strategies in quantitative trait loci mapping: conditioning on the genetic model. *Genetics* **148**, 525-535 (1998).
13. Visscher, P.M., Thompson, R. & Haley, C.S. Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013-1020 (1996).
14. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508 (2014).
15. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics* **98**, 1114-1129 (2016).
16. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics* **3**, e114 (2007).
17. The International Multiple Sclerosis Genetics Consortium. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics* **45**, 1353-1360 (2013).
18. Chen, W. *et al*. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719-736 (2015).
19. Benner, C. *et al*. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).
20. Manolio, T.A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
21. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
22. Nica, A.C. *et al*. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895 (2010).
23. Ongen, H. *et al*. Estimating the causal tissues for complex traits and diseases. *bioRxiv*, 074682 (2016).
24. Giambartolomei, C. *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
25. Sharma, G. *et al*. GPER deficiency in male mice results in insulin resistance, dyslipidemia, and a proinflammatory state. *Endocrinology* **154**, 4136-4145 (2013).
26. Meyre, D. *et al*. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature genetics* **41**, 157-159 (2009).
27. Jelinek, D., Heidenreich, R.A., Erickson, R.P. & Garver, W.S. Decreased Npc1 gene dosage in mice is associated with weight gain. *Obesity* **18**, 1457-1459 (2010).
28. Jelinek, D. *et al*. Npc1 haploinsufficiency promotes weight gain and metabolic features associated with insulin resistance. *Human molecular genetics* **20**, 312-321 (2010).
29. Bambace, C., Dahlman, I., Arner, P. & Kulyté, A. NPC1 in human white adipose tissue and obesity. *BMC Endocrine disorders* **13**, 5 (2013).

30. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
31. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience* **19**, 1442-1453 (2016).
32. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99**, 1245-1260 (2016).

Figure 1: eQTL discovery with different genotyping technologies. Number of autosomal eQTLs discovered in each tissue when genotype information is provided by arrays imputed into a reference panel (genotype) and by whole genome sequencing (sequence). There is a modest (3.7%) increase in the total number of eQTLs discovered with WGS over all tissues.

Figure 2: Relative enrichment of eQTLs discovered with different genotyping technologies in functional regions. Odds ratio and P value for enrichment (Fisher's Exact Test) of lead eQTL variants (LEVs) called from sequence being located in DNase hypersensitivity sites[11] relative to LEVs called from array derived genotypes. A total of 31 experiments related to the tissue from which RNA-seq was collected were analysed. The code given relates to the Roadmap Epigenomics code, Supplementary Table 2 lists the original experiment. All but enrichment of skin eQTLs in DHS assayed in NHDF-Ad Adult Dermal Fibroblast Primary Cells were Bonferroni significant (two tailed Fisher test, $P < 0.05$).

Figure 3: Distribution of the CaVEMaN estimated causal probabilities for lead eQTL variants (LEVs). We show the distribution of causal probabilities for all LEVs discovered in each of the tissues.

Figure 4: Probability of a lead eQTL variant (LEV) falling into a DHS region against causal probability. LEVs were divided into ten equally sized groups by causal probability and the proportion falling into DHS regions calculated for each group and each experiment. The complete line represents the median result across experiments; where there is more than one experiment for a given tissue, the dotted lines give the maximum and minimum across

experiments. We observe a linear relationship between the two probabilities. A full list of experiments can be found in Supplementary Table 2.

Figure 5: Proportion of functional variants falling into regions identified by single ChIP-seq experiments. We estimated the proportion of LEVs with causal probability of 1 falling into functional regions defined by the Roadmap Epigenomics Consortium by extrapolating from the relationship observed in Figure 4. Blood shows the highest proportions falling in annotations, for all other tissues we estimate only a minority of causal variants are in DHS regions.

Figure 6: High confidence causal variants (HCCVs) statistically associated with GWAS traits. Numbers of Bonferroni significant associations between HCCVs (causal probability > 0.8) and GWAS traits, divided by tissue type. HCCVs show more statistical associations with GWAS traits than other eQTLs, as a co-segregating signal will not be weakened by imperfectly captured markers.

Online Methods

TwinsUK data

Expression and genotype data from arrays.

RPKM expression quantifications and array genotype data used in this paper have been previously analysed[5,6] and the production of these data is described in full in Supplementary note 1.

Genotypes called from sequencing.

The vcf files, produced by the UK10K consortium[7], were downloaded from the European Genome-phenome Archive. When one monozygotic twin in the sample had been sequenced, the same data was used for the genetically identical sibling. Of 856 individuals with

expression, 552 have available sequence data (246 individuals had expression quantified in whole blood, 505 in adipose tissue, 523 in LCLs and 471 in skin). For multiallelic variants, dosage was calculated as 2 times the number of copies of the most common allele. Variants were filtered if the major allele had a frequency >0.99.

Ethics statement.

The St. Thomas' Research Ethics Committee (REC) approved on 20 September 2007 the protocol for dissemination of data, including DNA, with REC reference number RE04/015. On 12 March 2008, the REC confirmed this approval extended to expression data. Volunteers gave informed consent and signed an approved consent form before the biopsy procedure. Volunteers were supplied with an appropriate detailed information sheet regarding the research project and biopsy procedure by post before attending for the biopsy. Consent to link the RNA-seq data with the whole genome sequence data was approved by the TwinsUK Resource Executive Committee (TREC) on 22nd April 2015.

Geuvadis data.

BAM files for RNA-seq were downloaded from EBI ArrayExpress, accession code E-GEUV-3. These were mapped to the GRCh37 reference genome[33] using GEM version 1.7.1[34] and protein coding and lincRNAs were quantified using the GENCODE v19 annotation[35]. Population group was regressed out of RPKM values using a linear model, values were centred and scaled to mean 0, variance 1, and 50 principal components were removed. Genotype vcf files from phase 3 of the 1000 Genomes project[36] were downloaded from the 1000 Genomes website. In non-pseudo autosomal regions of the X chromosome, male dosage was calculated as twice the number of copies of the alternate allele. A minor allele frequency cut off of 0.01 was applied.

eQTL mapping.

eQTLs were mapped using fastQTL which tests for association between expression and genotype with a two-tailed Wald test[37]. To discover multiple independent eQTLs, a stepwise regression procedure was applied. Firstly, for each tissue, fastQTL was run with 10,000 permutations to discover a set of eGenes (FDR <0.01). Then, the maximum beta-adjusted P value (correcting for multiple testing across SNPs) over these genes was taken as the gene-level threshold. The next stage proceeded iteratively for each gene. At each iteration a cis scan of the window was performed, using 10,000 permutations and correcting for all previously discovered SNPs. If the beta adjusted P value for the LEV was not significant at the gene-level threshold, the procedure moved on to the backward step. If this P value was significant, the LEV was added to the list of discovered eQTLs as an independent signal and the forward step proceeded to the next iteration.

Once the forward stage was complete for a given gene, a list of associated SNPs was produced which we refer to as forward signals. The backwards stage consisted of testing each forward signal separately, controlling for all other discovered signals. For each forward signal we ran a cis scan over all variants in the window using fastQTL, fitting all other discovered signals as covariates. If no SNP is significant at the gene-level threshold the signal being tested is dropped, otherwise the LEV from the scan was chosen as the variant that represented the signal best in the full model.

Properties of LEVs estimated using sequence and genotyping arrays.

We investigated the differences between LEVs identified using sequence data and data from genotyping arrays to better understand the slight increase in power we see using sequence. The minor allele frequency (MAF) of eQTLs called using sequence is slightly lower than those

identified using genotype data (median MAF of 26.0% compared to 27.4%, two tailed Mann-Whitney U test $P=5.52\times10^{-21}$, Supplementary Figure 2). We find that 3,383 out of 22,656 LEVs called using sequence are removed from the array data due to INFO score less than 0.8, the majority of these failed imputation criteria based on the HumanHap300 array (3,334 failed on this array, 2,290 failed on the HumanHap610Q, and 2,241 failed on both, Supplementary Figure 3). Finally, for remaining 19,273 sequence LEVs where the genotype imputation passed the quality filters, we see good agreement between calls made using the two technologies, with a median proportion of different calls of only 0.94%. However, there are a small minority of LEVs (0.93%) where we see a larger discrepancy between the two call sets, with more than 10% of individuals showing differences. Together, these results suggest that both genotyping arrays with more SNPs and larger reference panels which enumerate more haplotypes will further reduce the power differences between studies using sequencing and those using genotyping arrays.

Enrichment analysis.

Bed files listing DNase hypersensitivity sites, produced by the Roadmap Epigenomics consortium[38], were downloaded from the NCBI ftp site. Experiments were linked to tissues for which RNA-seq was available using Supplementary Table 2. Over each ChIP-seq RNA-seq combination, the odds ratio for enrichment was calculated using the number of LEVs called using sequence and the number of LEVs called using array-based genotypes falling within regions called in the experiment and the total numbers of eQTLs. A two tailed Fisher's Exact test was performed to test the hypothesis that equal proportions of sequence and genotype LEVs fell in these regions.

Simulations.

For all discovered eQTLs, the LEV for association was identified and its minor allele frequency and distance to the transcription start site calculated. Beta and sigma coefficients from a regression of expression on the LEV were also estimated. Then a matched SNP was chosen, with a distance to transcription start site of a gene within 1 kb of the original, and minor allele frequency within 0.025. Simulated expression was produced by multiplying SNP genotype by beta and adding a random normally distributed term with standard error of sigma. Five simulated datasets were produced for each TwinsUK tissue, eQTL mapping was applied to each looking only for primary eQTLs, and the rank of the nominal P value for the causal variant was collected.

CaVEMaN.

A frequentist definition of a causal probability.

A number of methods have been proposed using Bayesian methodologies to estimate the probability that a variant is causal for an effect on expression, combining prior distributions with likelihoods to estimate posterior probabilities[15,39,40]. We, however, use a frequentist definition of the probability of being causal. Causal probabilities are assigned to LEVs with the following property: if an eQTL is sampled randomly from the set of all eQTLs that have a causal probability equal to a number x, the probability that a causal variant is chosen is equal to x. In this way it matches an intuitive understanding of what a causal probability is: if a LEV is chosen at random, the probability that a causal variant will be chosen is equal to the estimate from CaVEMaN.

Learning parameter estimates from simulations.

Firstly, we used the simulations where a specific variant has been chosen to act as the causal variant to estimate the probability the causal variant would be the i^{th} ranked SNP in an eQTL mapping. This is done by calculating the proportion of times this occurred across all tissues and simulations (this quantity is denoted p_i, Supplementary Figure 1). As CaVEMaN focuses on the top 10 ranked variants from an eQTL analysis, p_i, i from 1 to 10, were normalised to sum up to 1.

Multiple variants affecting expression of one gene.

Previous fine-mapping approaches can be categorised into two classes: those that assume only one genetic signal affects the phenotype[39], and those which map multiple genetic signals simultaneously[15,40]. CaVEMaN takes a different approach, in that the procedure is separated into two steps: firstly, a stepwise regression approach is used to estimate the number of eQTL affecting the expression of the gene, and then each independent eQTL is mapped separately. The advantage of this is there exists a well-grounded statistical methodology for answering the question of multiple independent variables affecting expression which deals with issues of multiple testing and significance.

Once a set of eGenes and the independent eQTLs affecting them has been identified, we create new "single signal" expression phenotypes. For each eQTL these are made by regressing out all other eQTLs discovered for the gene, producing an expression phenotype which reflects the signal from only one eQTL.

Calculating CaVEMaN score.

This new matrix of expression data was sampled with replacement 10,000 times to create 10,000 new datasets of the same size. A cis eQTL mapping testing association using a two

tailed test for significant correlation was run on each of these datasets, and the proportion of times a given SNP was ranked *i*, *i* from 1 to 10 was calculated (denoted by F_i, this is an estimate of the probability that SNP would be the rank i^{th} most associated SNP). The CaVEMaN score was defined as $\sum_{i=1}^{10} p_i F_i$, i.e., the sum of the product of the probability the SNP is ranked *i* in an eQTL analysis with the probability the i^{th} ranked SNP is causal for the association.

Calibrating CaVEMaN score for LEVs using simulations.

Finally, we further exploited the simulations to calibrate the CaVEMaN score of the LEV. CaVEMaN was run on all simulated data. Then, across all simulated datasets (removing blood as this was an outlier resulting in less conservative estimates of causal probabilities) we divided the CaVEMaN scores of the LEVs into twenty quantiles. Within each quantile, we calculated the proportion of times the lead SNP was the causal SNP and then drew a monotonically increasing smooth spline from the origin, through the 20 quantiles, to the point (1, 1) using the gsl interpolate functions with the steffen method (gsl-2.1, Supplementary Figure 4). This function provides our mapping of CaVEMaN score of the lead SNP onto causal probabilities, and we applied this function to the CaVEMaN scores of the LEV to estimate their causal probabilities.

Validating the method with simulations in the Geuvadis data.

The CaVEMaN method uses parameters estimated from simulations based on the UK10K expression data (chiefly the distribution of ranks of causal eQTLs and the relationship between CaVEMaN score and causal probability), meaning that these simulations cannot later be used to validate the CaVEMaN estimates. We have run further simulations using the Geuvadis data to demonstrate that the estimates of the causal probability for the LEVs are well calibrated,

when parameters are estimated from separately from the dataset being analysed. A total of five simulations were run, again using effect size and residual variance estimated from the original data. In Supplementary Figure 5 we plot binned estimates of the estimated causal probabilities against the proportion of times the LEV is the causal variant. We see good agreement between our estimates and the true causal probabilities for these bins: the minimum, median and maximum difference between the estimate and the true values are 0.0056, 0.036 and 0.071 respectively.

In addition to this, we have run a further simulation to test the behaviour of the model when there are weaker eQTL effects which are not detected by the original multiple eQTL mapping strategy. As before, we simulate a primary eQTL with minor allele frequency, effect size and distance to the transcription start site matched to an eQTL discovered in the original analysis. Then, a second variant is chosen randomly in the cis window, with minor allele frequency greater than 0.05, and we use this variant to simulate an extra eQTL effect on the phenotype, with effect size one half of the primary eQTL. Then a residual noise term was generated such that the primary eQTL explains the same proportion of variance as the original matched eQTL. In Supplementary Figure 5 we see that there is still good agreement when estimating the causal probabilities of the primary eQTL to the known ground truth.

Comparing results on TwinsUK LCL data with results from CAVIAR.

CAVIAR, along with equivalent Bayesian methods[16-19], have previously been suggested as fine-mapping methods for estimating credible sets of SNPs with a given probability of containing the causal variant. For genes with an eQTL in LCLs, we applied CAVIAR[14] to produce another estimate of causal variant probability for comparison. As CAVIAR is limited in the number of SNPs it can analyse, we first extracted all variants with $P < 0.01$, up to the first 50. The Z scores

for these variants were produced, with the correlation matrix of these SNPs, and CAVIAR was run with the default settings. There was good agreement on the causal probabilities of the LEV (spearman $\rho = 0.856$, $P < 10^{-216}$, Supplementary Figure 6), but the CAVIAR method produced more conservative estimates of the causal probabilities (median probability 0.12 vs 0.29). As the CaVEMaN estimates are calibrated using simulations, this suggests that the CAVIAR estimates are on average underestimates of the true probabilities, which could be due to a combination the priors not reflecting the true regulatory landscape and the sample size being insufficient to overcome this. CAVIAR does not suggest adjusting the priors when studying expression rather than GWAS trait associations, despite the fundamentally different genetic architectures and sample sizes between these types of studies. The approach of calibrating estimates of probabilities using simulations could also be easily extended to other fine mapping methods such as CAVIAR.

Comparison of simulation results with those produced using dap-g.

We have compared the results of CaVEMaN when applied to one of the simulation datasets to results produced using dap-g[15], a recent method proposed for fine-mapping. For each simulated gene expression, all SNPs in the cis window were extracted and dap-g was run, specifying the option -ld_control 0.25. Then, for a comparable estimate of the posterior probability of the LEV, we extracted the highest posterior probability of any single variant model, and conditioned this on only one genetic signal by dividing this by the sum of the posterior probabilities of all single SNP models. The two methods identify exactly the same sets of LEVs and there is good agreement between the estimates of causal probabilities (Spearman $\rho = 0.95$, $P<10^{-216}$). However, plotting the causal probabilities against the

proportion of LEVs which are the causal variants, we see that dap-g underestimates this quantity (Supplementary Figure 5).

Application of simulations to other datasets.

The Geuvadis dataset differs in many aspects from the TwinsUK data that the CaVEMaN method was trained on: Geuvadis samples were sequenced in multiple laboratories rather than just one, Geuvadis uses a multi-ethnic cohort implying a different linkage structure in the genome, a different mapper (in particular, a splice-aware mapper) was used to quantify the data, and the tissue type, sample size and ability to map eQTLs are all different from three out of four TwinsUK tissues. This shows the parameters estimated in TwinsUK are robust to a range of factors. However, in the future, similar datasets with thousands of samples are expected and it is possible that our proposed method will not generalise to this case. For this reason, we provide methods to repeat these simulations in new datasets, described on the accompanying website.

Statistical associations between eQTLs and GWAS traits from summary statistics.

We have downloaded the GWAS summary statistics for 16 different GWAS traits: autism[41], birth weight[42], body mass index (analysing all ancestries)[43], coronary artery disease[44], Crohn's disease[45], diabetes[46], fasting glucose[47], fasting insulin[47], height[48], high-density lipoprotein[49], irritable bowel disease[45,50], low-density lipoprotein[49], schizophrenia[50,51], total cholesterol[49], triglycerides[49], and ulcerative colitis[45]. Data on birth weight trait has been contributed by the EGG Consortium using the UK Biobank Resource and has been downloaded from www.egg-consortium.org. For all LEVs, the P value for each trait was extracted (if available) and the qvalue package[21] was used to estimate π_1, the proportion of alternate hypotheses (i.e.,

association between variant and GWAS trait). Finally, Bonferroni significant GWAS associations for HCCVs were reported, controlling for multiple testing across all phenotypes and variants.

Testing HCCVs associated with GWAS traits for co-segregation using coloc.

For HCCVs significantly associated with GWAS traits, we used the coloc method[24] to test the hypothesis of shared causal mechanism. P values for association, available for both expression and GWAS associations, were extracted in a 200,000bp region around the eQTL. Minor allele frequencies for the variants were extracted from the 1000 Genomes Phase 3 release[36]. After running coloc, we report the probability of a shared causal variant for both associations, conditional on genuine associations existing for both traits (P(H3) / (P(H3) + P(H4)) reported by coloc).

Regulatory Trait Concordance method for testing for co-segregation with the NHGRI-EBI catalog GWAS associations.

We downloaded the NHGRI-EBI Catalog of reported genome-wide significant associations from the EBI website on the September 2016 and removed all with $P>5\times10^{-8}$ and where the variant was not listed in dbSNP build 148[52], leaving 11,636 reported associations. RTC, as implemented in QTLtools [53], was applied with the default settings to look for sharing of these GWAS variants with eQTLs. As the RTC statistic is uniformly distributed under the null hypothesis of two separate causal loci independently located within the hotspot, 1 - RTC can be interpreted as a P value for a shared causal variant. The qvalue package[21] estimated π_1, the proportion of GWAS/eQTLs signals in the same recombination interval with the same causal variant.

Code availability.

Code for correcting the expression datasets for multiple eQTLs, running the CaVEMaN method, converting the CaVEMaN score to a causal probability and repeating simulations on new datasets can be found here:

https://github.com/funpopgen/CaVEMaN.

Data availability

BAM files for the RNA-seq are available from EBI ArrayExpress, accession code E-GEUV-3 (Geuvadis cohort) and the European Genome-Phenome Archive, study ID EGAS00001000805 (TwinsUK cohort). Whole genome sequence is available from the European Genome-Phenome Archive, study ID EGAS00001000108 (TwinsUK) and the 1000 Genomes website (Geuvadis).

References

5. Brown, A.A. et al. Genetic interactions affecting human gene expression identified by variance association mapping. Elife 3, e01381 (2014).
6. Buil, A. et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet 47, 88-91 (2015).
7. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. Nature 526, 82-90 (2015).
11. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330 (2015).
14. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. Genetics 198, 497-508 (2014).
15. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. The American Journal of Human Genetics 98, 1114-1129 (2016).
16. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS genetics 3, e114 (2007).
17. The International Multiple Sclerosis Genetics Consortium. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nature genetics 45, 1353-1360 (2013).
18. Chen, W. et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics 200, 719-736 (2015).
19. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493-1501 (2016).
21. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100, 9440-5 (2003).
24. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 10, e1004383 (2014).

33. Lander, E.S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).
34. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9, 1185-8 (2012).
35. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760-74 (2012).
36. Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68-74 (2015).
37. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics 32, 1479-85 (2016).
38. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317 (2015).
39. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS genetics 9, e1003486 (2013).
40. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS genetics 11, e1005176 (2015).
41. Robinson, E.B. et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nature genetics 48, 552 (2016).
42. Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. Nature 538, 248-252 (2016).
43. Locke, A.E. et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197 (2015).
44. Nikpay, M. et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics 47, 1121 (2015).
45. Liu, J.Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics 47, 979-986 (2015).
46. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. Nature 536, 41-47 (2016).
47. Manning, A.K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nature genetics 44, 659-669 (2012).
48. Wood, A.R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics 46, 1173-1186 (2014).
49. Consortium, G.L.G. Discovery and refinement of loci associated with lipid levels. Nature genetics 45, 1274-1283 (2013).
50. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421 (2014).
51. Dabney, A. & Storey, J.D. qvalue: Q-value estimation for false discovery rate control.. R package version 1.40.0.
52. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29, 308-11 (2001).
53. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. Nat Commun 8, 15452 (2017).